Multimodal control of sensors on multiple simulated unmanned vehicles.

Baber, Christopher; Morin, C; Parekh, Manish; Cahillane, M; Houghton, RJ

DOI:
10.1080/00140139.2011.597516

Document Version
Early version, also known as pre-print

Citation for published version (Harvard):
Baber, C, Morin, C, Parekh, M, Cahillane, M & Houghton, RJ 2011, 'Multimodal control of sensors on multiple simulated unmanned vehicles.', *Ergonomics*, vol. 54, no. 9, pp. 792-805.
https://doi.org/10.1080/00140139.2011.597516

Ergonomics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/terg20

Multimodal control of sensors on multiple simulated unmanned vehicles

C. Baber [a], C. Morin [b], M. Parekh [a], M. Cahillane [b] & R.J. Houghton [c]

[a] Electronic, Electrical and Computer Engineering, The University of Birmingham, Birmingham B15 2TT, UK

[b] Department of Informatics and Systems Engineering, Cranfield University, Shrivenham SN6 8LA, UK

[c] Horizon Digital Economy Research, Sir Colin Campbell Building, University of Nottingham, Nottingham NG7 2TU, UK

Available online: 25 Aug 2011

To cite this article: C. Baber, C. Morin, M. Parekh, M. Cahillane & R.J. Houghton (2011): Multimodal control of sensors on multiple simulated unmanned vehicles, Ergonomics, 54:9, 792-805

To link to this article: http://dx.doi.org/10.1080/00140139.2011.597516

Ergonomics
Vol. 54, No. 9, September 2011, 792–805

Taylor & Francis
Taylor & Francis Group

Multimodal control of sensors on multiple simulated unmanned vehicles

C. Baber[a]*, C. Morin[b], M. Parekh[a], M. Cahillane[b] and R.J. Houghton[c]

[a]*Electronic, Electrical and Computer Engineering, The University of Birmingham, Birmingham B15 2TT, UK;* [b]*Department of Informatics and Systems Engineering, Cranfield University, Shrivenham SN6 8LA, UK;* [c]*Horizon Digital Economy Research, Sir Colin Campbell Building, University of Nottingham, Nottingham NG7 2TU, UK*

(*Received 12 December 2010; final version received 10 June 2011*)

The use of multimodal (speech plus manual) control of the sensors on combinations of one, two, three or five simulated unmanned vehicles (UVs) is explored. Novice controllers of simulated UVs complete a series of target checking tasks. Two experiments compare speech and gamepad control for one, two, three or five UVs in a simulated environment. Increasing the number of UVs has an impact on subjective rating of workload (measured by NASA-Task Load Index), particularly when moving from one to three UVs. Objective measures of performance showed that the participants tended to issue fewer commands as the number of vehicles increased (when using the gamepad control), but, while performance with a single UV was superior to that of multiple UVs, there was little difference across two, three or five UVs. Participants with low spatial ability (measured by the Object Perspectives Test) showed an increase in time to respond to warnings when controlling five UVs. Combining speech with gamepad control of sensors on UVs leads to superior performance on a secondary (respond-to-warnings) task (implying a reduction in demand) and use of fewer commands on primary (move-sensors and classify-target) tasks (implying more efficient operation).

Statement of Relevance: Benefits of multimodal control for unmanned vehicles are demonstrated. When controlling sensors on multiple UVs, participants with low spatial orientation scores have problems. It is proposed that the findings of these studies have implications for selection of UV operators and suggests that future UV workstations could benefit from multimodal control.

Keywords: unmanned vehicles; multimodal interaction; speech recognition

Introduction

Developments in unmanned (uninhabited) vehicles (UVs) have increased their capability for autonomy (Finn and Scheding 2010). Thus, unmanned air vehicles (UAVs) can fly with little direct intervention; rather the operator defines way-points to which the vehicle routes itself. Of course, there remain UVs that require the direct control of a pilot or drivers, such as larger airborne vehicles or ground-based vehicles, but in these either a second operator takes responsibility for managing the sensors and analysing the imagery, or there is separation of the vehicle manoeuvring tasks from effector or sensor control tasks. In this paper, attention is focused on interaction with autonomous UVs (either air or ground), leaving the operator with the tasks of managing sensors and interpreting sensor data.

The human factors issues of interacting with UVs have been comprehensively reviewed by Chen *et al.* (2007). Much of the research focuses on the control of the vehicle rather than the sensors or effectors that it carries. Thus, this paper will contribute to the debate on human factors of UVs by considering the operator's interaction with sensors on the vehicles.

Multimodal interaction with UVs

There is a long tradition of combining speech with manual control to effect multimodal control of computer systems (e.g. Bolt 1980, 1984, Martin 1989, Schmandt *et al.* 1990, Biermann *et al.* 1992, Thorisson *et al.* 1992, Hauptmann and McAvinney 1993, Baber 1997, Cohen *et al.* 1997, 2002, Oviatt *et al.* 1997, 2000, Baber and Mellor 2001).

Multimodal display appears to both reduce UV operator workload and provide access to multiple streams of information (Dixon and Wickens 2003, Wickens *et al.* 2003, Trouvain and Schlick 2007, Maza *et al.* 2009). Auditory presentation of information can

*Corresponding author. Email: c.baber@bham.ac.uk

ISSN 0014-0139 print/ISSN 1366-5847 online

http://dx.doi.org/10.1080/00140139.2011.597516
http://www.tandfonline.com

be combined with ongoing visual tasks (Helleberg *et al*. 2003), and these improvements can be particularly important when dealing with multiple UVs, provided that they do not interfere with other auditory warnings (Donmez *et al*. 2009). However, combining the control of a UV with other tasks can impair performance on target detection (Dixon and Wickens 2003, Chen 2008) and reduce situation awareness (Luck *et al*. 2006). Draper *et al*. (2003) compared speech and manual data entry when participants had to manually control a UV and found speech yielded less interference with the manual control task than manual data entry. Chen (2008) showed that target detection was significantly impaired when participants had to combine search with control of the vehicle, in comparison with a condition in which the vehicle was semi-autonomous. Moreover, individuals with higher spatial capability performed target detection tasks better than people with lower spatial capability (Chen 2009).

In general, these applications use manual control for manipulating objects and spoken response for issuing commands. From this perspective, the benefit to be obtained from multimodality is the division of activity between two response modalities. Experimental studies suggest that it is possible to 'share' attention between modalities. In the multiple resources tradition (Wickens and Liu 1988), differences in performance when using speech or manual control were attributed to the compatibility between response and processing 'codes'. Thus, for example, a task that involved detecting objects in space would require a spatial processing code, which could be best responded to manually. Similarly, a task that involved detecting auditory events would require an auditory code, which could best be responded to vocally. Furthermore, as there was assumed to be 'spare' capacity in the 'unused' codes, manual control of one task could be performed simultaneously with spoken response in another task (Wickens *et al*. 1983, 2003, Wickens and Liu 1988, Wickens and McCarley 2008). This multiple resource description contrasts with the single-channel hypothesis (Kahneman 1973), which sees no obvious distinction in processing codes. Rather, there is a central executive, which coordinates the allocation of 'resource' to tasks regardless of 'code'. While the multiple resource model drew inspiration from the dual-code approach that was at the heart of Baddeley and Hitch's (1974) notion of working memory, recent developments of this notion have emphasised the importance of the central executive (Baddeley 1998). Thus, contemporary explanations of dual-task performance relies less on the cooperation or competition between discrete processing codes and more on the scheduling of tasks. For the purposes of this paper, the question of task performance becomes one of managing the competing demands of the primary control and classification tasks with those of the competing 'respond-to-warnings' secondary task. It is proposed that differences in performance could arise partly from the compatibility between the spatial tasks involved in controlling the sensors and classifying targets and partly from the ability to schedule task demands, particularly as the number of UVs increase.

Interaction with multiple UVs

Taylor (2006) notes that increasing autonomy allows multiple UVs to fly with little or no manual intervention from the human operator. However, this does not mean that the human will be removed from the control loop. Rather, the tasks of the operator will shift towards supervisory control of the vehicles and analysis of the information from the sensors on the vehicle. Taylor suggests that four UVs per operator is a typical design aim for future systems. Cummings *et al*. (2007) present a meta-review of UVs with different levels of automation and suggest that studies converge on four to five vehicles when control and decision making are performed by the operator. Their models of operator performance suggests that control of one UV is superior to two, three or four (which result in similar performance) with degradation in performance when controlling five or more UVs. Liu *et al*. (2009) found significant impairment in performance on secondary tasks (response to warnings or status indicators) when operators controlled four UVs, in comparison with controlling one or two UVs. Thus, one can propose some difference in performance when the number of UVs being controlled changes from one to two and a further change when the number increases from two to four or more, and that these differences in performance could be reflected through variation in performance on secondary tasks.

Rationale for this work

As UVs get smaller so control will move from rear echelons to front-line troops (with UVs that can be carried and operated by one- or two-man crews) and there will be increasing challenges to the design of user interfaces to support the combination of these control activities with other demands on the operator. A multimodal interface might allow the operator to distribute attention between control of the UV, interaction with sensors and other demands from the local environment. From the discussion of multimodal interaction, it is proposed that participants will seek to allocate functions to speech and manual controls, if available. It is also proposed that performance using

speech alone will be inferior to that of the other conditions because of the incompatibility between speech and the spatial tasks required in managing the sensors or searching for targets. Furthermore, it is suggested that the scheduling of tasks will be a key indicator of performance, and this should lead to impairment of the secondary task.

Second, as UVs get smaller then it is likely that a single operator will be responsible for more than one vehicle as they could be flown in 'swarms' or small formations. From the discussion relating to the control of multiple UVs, it is suggested that performance should be impaired when controlling 5 or more UVs, in comparison with controlling fewer UVs.

Experiment 1

Aims

The aim of experiment 1 is to compare the performance of participants when using single and multiple modalities to control sensors on simulated UVs. Missions will be performed using just speech recognition or just gamepad control or a combination of the two. As commands can be performed using either modality, the third (multimodal) condition allows participants discretion as to which modality they chose for a given task. Thus, the objectives of this experiment are to explore the impact of speech and manual interaction on the performance of sensor control tasks and to consider how users might choose to allocate functions to each modality.

Design

One can assume that future systems will be autonomous and that the role of the operator will be to manage the sensors. Thus, the first assumption for the design of the multimodal UV test-bed used in this study is that control will focus on sensors. It was further assumed that, even if the UV was autonomous, the operator would need to deal with warnings concerning vehicle state. These two assumptions define the control tasks of the operator:

- The primary tasks concern the control of the sensors and the classification of targets[1]. The control actions (described in detail below) involve issuing commands to the sensors on the UV (to move these left or right to search for targets) and the classification of targets involves indicating the type of target detected.
- The secondary tasks concerned the need to respond to 'system warnings'. This involves a response to a stimulus (described below) and performance is defined in terms of response time.

Participants

Altogether, 16 undergraduate students of Computer Interactive Systems in the School of Electronic, Electrical and Computer Engineering at The University of Birmingham participated in this experiment. Their mean age was 23 years and there were three female and 13 male students. All participants were naive to the task of UV operation and target detection. It is felt that, for these experiments, this is not a significant issue. After all, the control of multiple UVs is not a common operational requirement at present and the use of speech recognition for UV control is not standard operationally, so it is unlikely that participants could be recruited with experience of these systems. The participants had experience of playing video games and were familiar with gamepad controls, but not speech recognition. It was felt that this would be a reasonable reflection of the abilities of operators as speech recognition is still fairly unfamiliar to most people.

Equipment

In terms of interaction devices, discussion with subject matter experts suggests that some manufacturers are migrating their control devices towards those used on games consoles. The justification for this seems to be a combination of familiarity of operators with these types of control and a well-defined button-set that can be exploited for issuing commands. Consequently, the Multimodal UV Test-bed uses a gamepad control for manual control of the vehicles (see Figure 1). For this study, the control of the UV, and identification of targets, involved a common set of 11 commands[2]. These commands were issued using controls either the gamepad or Microsoft's SAPI speech recognition application (Microsoft Corporation, Redmond, WA, USA). In order to ensure parity across controls, any command that involved movement of the sensors was defined in terms of a set number of pixels (which means that slewing of the sensors would involve more than one button press or spoken command).

The user interface for the Multimodal UV Test-bed was written in Microsoft .Net C# (Microsoft Corporation). The screen is divided into three main sections, as shown in Figure 2. The left-hand side of the screen shows a map with the routes that the UVs will follow; the right-hand side of the screen shows the imagery from the sensors; the bottom of the screen

Figure 1. Microsoft Xbox Control with multimodal UxV test-bed.

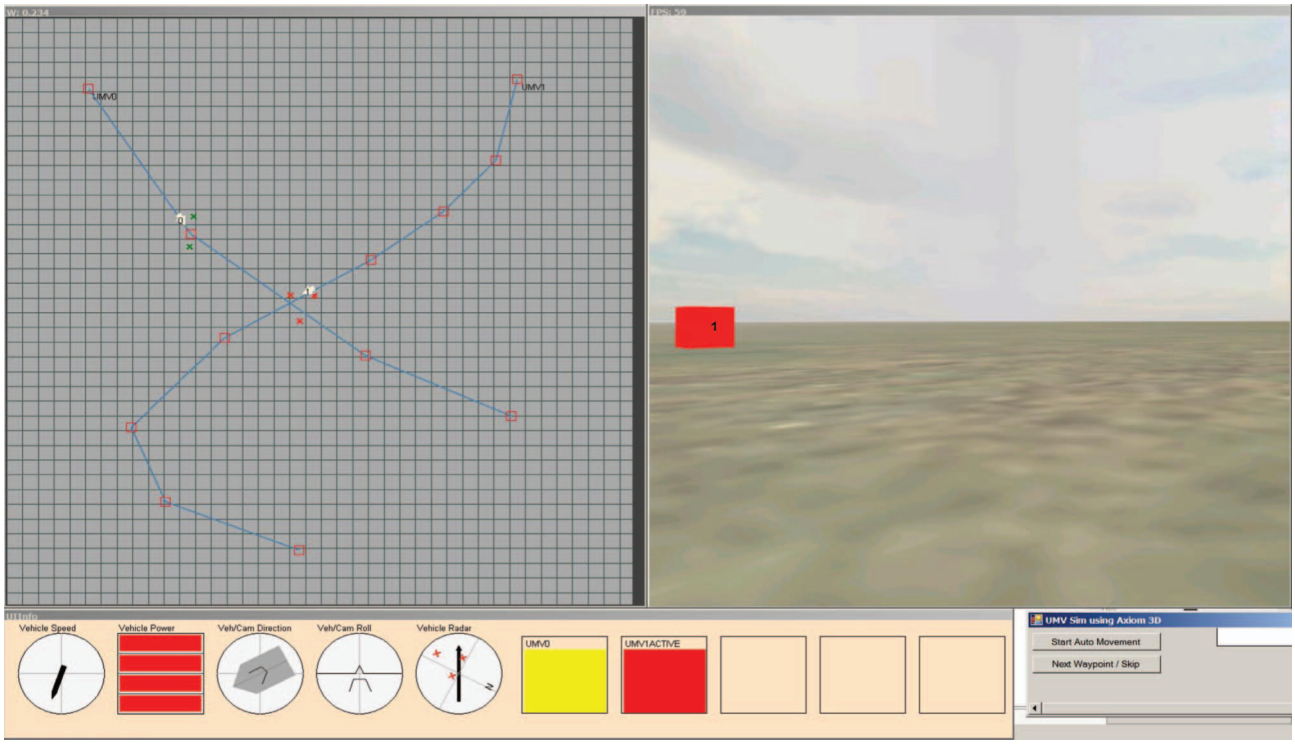

Figure 2. User interface.

shows system status for specific UVs and which UV was currently being controlled.

For sensor imagery, a simple block-world was created (see Figure 2). The targets were cubes with a number on one face. This provided sufficient complexity to allow simple decisions to be made by participants, i.e. 'is the number on the cube odd or even?' The view from the UV reflected the type of vehicle, such that the unmanned ground vehicle (UGV) would view an object located in the horizontal plane

on the ground and the UAV would view an object located in the vertical plane. The different vehicles required different levels of 'zoom' to read the numbers on their sides.

For the UV routes, the application was run in 'plan mode'. In plan mode, the experimenter can select (from a menu that pops up from the bottom left of the screen) different types of UV and each UV results in different imagery from the sensors. Once a UV is selected, the experimenter can define a number of way-points on the map to define a route (in planning mode) and position 'targets' around the route. Each 'target' can be allocated an identification code (see Figure 3).

Dependent variables

All commands that the participant issued were logged in terms of time and in terms of proximity to target. The primary performance measures involve:

- Number of commands issued using the different modalities. The number of commands issued using speech is hypothesised to be less than for the gamepad. This could arise from the spatial nature of the sensor manipulation (which would favour the gamepad) and also from the categorical nature of classification (which could favour speech).
- The distance to the target at identification was calculated in 2-D space as the number of pixels from the target to the front of the UV. As the speed of the UVs is known, this allows the appearance of targets to be scheduled during experimental design. In each trial, participants had to attend to six targets (the number of targets 'seen' by a single UV, therefore, varies with the number of UVs in the trial). The appearance of targets was scheduled so that they did not overlap temporally; the relative ordering of allocation of target between the UVs is pseudo-random: as a UV approaches the location of a target, a 'radar' display in the bottom of the window showed the position of the target relative to the UV and a proximity warning was presented.
- The time to respond to system warnings. At various points in each run, a set of warnings was presented. The warnings were timed to occur when the UV was not in close proximity to the target. This was intended to reduce conflict in task demands on the participant. In each run there were six warnings, to which the participant needed to respond, for each UV. There was a total of six warnings per UV for a total of 12 per trial. These were again pseudo-randomly distributed along a rectangular distribution (i.e. each target appeared within a separate 25 s window; 5 min = 300 s / 12 = 25 s). Response to warnings would depend on the manner in which the different tasks are scheduled and is expected to provide an index of task difficulty and workload.

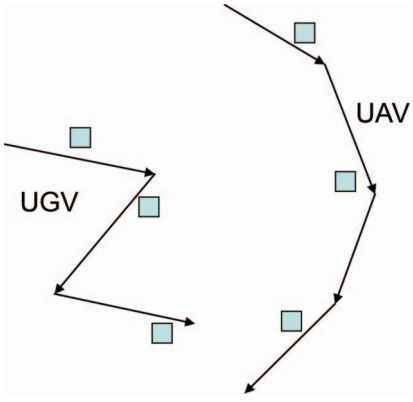

Figure 3. Schematic of routes and targets. UGV = unmanned ground vehicle; UAV = unmanned air vehicle.

In terms of baseline for these measures, it assumed that the higher the distance to target score, the better the performance, because the participant is able to make a response earlier in the decision cycle; the lower the time to respond to warnings the better, because the participant is able to deal with system problems quickly in order to return to the primary task.

Subjective measures

In addition to objective measures of task performance, the studies employ two forms of subjective measure. The first subjective measure is a questionnaire that is designed to elicit the participants' spatial capability and the second is the NASA-Task Load Index (TLX) subjective workload scale.

(i) Spatial-orientation was measured using Hegarty and Waller's (2004) revised version of the Object Perspective Test (Kozhevnikov and Hegarty 2001)[3]. For each item in this paper-and-pencil test, the same array of seven objects was displayed on the top half of an A4 sheet. Participants were required to imagine standing at the position of an object within the array (the station point) and looking at another object. At the bottom of the page a circle was displayed with the

imagined position (station point) indicated in the centre and the imagined perspective depicted by a vertical arrow pointing upwards. The task was to indicate the direction to a third object from their imagined heading, referred to as the target object, by drawing an arrow from the centre of the circle pointing towards the target object (see Figure 4). The score for each item was the absolute deviation (°) between the participant's response and the correct direction to the target. As the test measures an error score, scores were linearly transformed by subtracting the average error score from 180° so that higher scores corresponded to better performance. The revised Object Perspective Test has a reliability of 0.79 and 0.85, as measured by the Cronbach alpha statistic (Hegarty and Waller 2004).

(ii) Subject workload was measured using the NASA-TLX (Hart and Staveland 1988). This uses six dimensions to assess subjective workload: mental demand; physical demand; temporal demand; performance; effort; frustration. To obtain ratings for these dimensions, 20-step bipolar scales are used. A weighting procedure is used to combine the six individual scale ratings into a global score. Originally, this procedure required a paired comparison task to be performed prior to the workload assessments. However, several studies have shown that using paired-comparisons for scaling does not affect results (Byers *et al.* 1989, Nygren 1991, Moroney *et al.* 1995). Thus, the unweighted version of the scale was administered for these studies. Participants were asked to indicate their perceived workload at the end of each block of trials under each experimental condition.

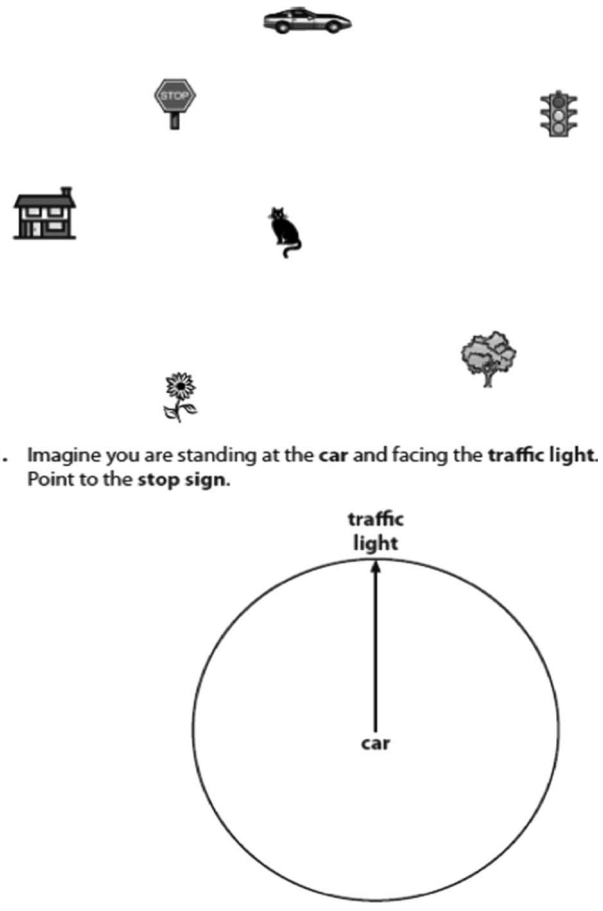

Figure 4. Example item from the Objective Perspective Test.

Procedure

For experiment 1, following completion of the spatial orientation task, participants were presented with a short (5 min) demonstration of the application, being controlled using either speech or gamepad, and given an opportunity to complete a practice run. The practice run took 5 min to complete and was, in the first instance, controlled using either speech or gamepad (depending on which condition participants were allocated). Following the trial run, participants performed three runs under the condition that they had practised. The trial runs involved two UVs (one UAV and one UGV) each with six warnings and three targets. Each trial run took 5 min. Following the set of three trial runs, participants received a demonstration of the control that they had not used. This was followed by a practice run using this control and then a set of three trial runs using this control. The two control conditions were counter-balanced across participants. Following performance using single modalities, participants were given an opportunity to practise using both modalities and then asked to complete three trials in this multimodal condition.

Each command was logged in terms of time and distance to target. For each trial run, the average of these data was used for analysis. Following each block of trials, participants completed the NASA-TLX to measure subjective workload. Analysis was conducted using three-way repeated measures ANOVA, which involved trial (1–3) × modality (speech, gamepad, multimodal) × vehicle (UAV × UGV).

Results

First, the analyses of results across all three trials are compared and then the results of the third trial (which allowed participants choice of modality) are considered.

Commands issued

There is a significant main effect of modality on the number of commands issued [$F(2,288) = 100.229$, $p < 0.0001$], but no other effects. Figure 5 shows that this effect can be explained primarily by the lower number of commands being issued when using the speech modality.

Distance to target

There is a significant main effect of vehicle for distance to target [$F(2, 288) = 472.576$, $p < 0.0001$] but no other. Figure 6 shows that UVs tend to be closer to the target in the 'air' mode than the 'ground' mode.

Time to respond to warning

There is a significant main effect of modality for time to respond to warnings [$F(2,288) = 6.354$, $p = 0.002$] but no other effects. Figure 7 suggests that the slowest responses occur when using speech response to targets in the 'air' condition and that the fastest responses occur in the multimodal conditions (for either ground or air conditions).

Subjective workload

Figure 8 indicates that there are no differences between conditions in terms of subjective workload rating across the different control modalities.

Influence of spatial ability

Performance on the spatial test was used to divide the participants into two groups. The mean score was used to define the split and this gave 10 participants with scores less than the mean and seven participants with scores greater than the mean. While this is a fairly crude break-down, it does create two groups for further analysis. Comparison of the results was performed using the Mann-Whitney U-test (because the data were assumed to be non-parametric). There were no significant differences between the conditions.

Distribution of commands in the multimodal trial

The third condition showed significant main effects of modality on command [$F(2,96) = 25.791$, $p < 0.0001$] and response time [$F(2,96) = 8.486$, $p = 0.004$]. There was no effect of modality on distance to target but a main effect of vehicle type [$F(1,96) = 161.487$, $p < 0.0001$]. These results reflect those of the initial trials. In order to explore the main effect of modality on command, the commands that people issued using

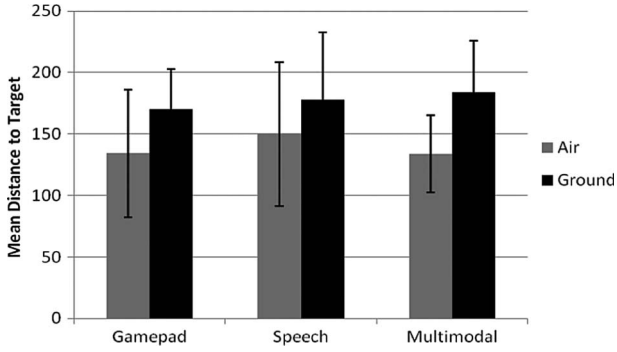

Figure 6. Mean distance to target, averaged over all trials for 'air' and 'ground' vehicles' time to respond to warnings.

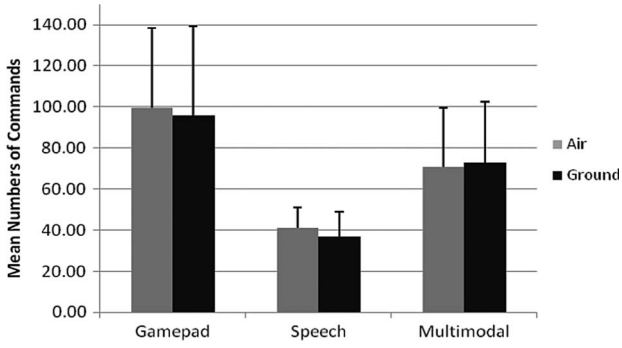

Figure 5. Mean number of commands for each modality averaged over all trials for 'air' and 'ground' vehicles.

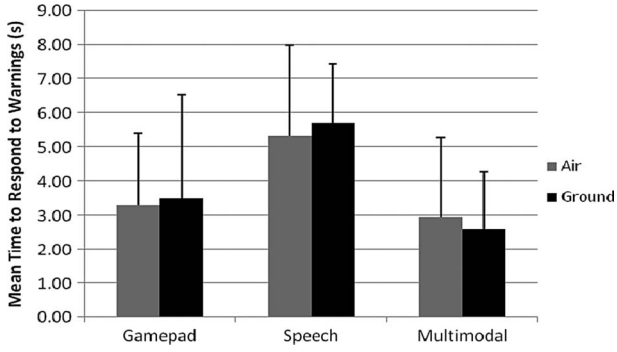

Figure 7. Mean time to react to warnings for each modality in the first, second and third trial.

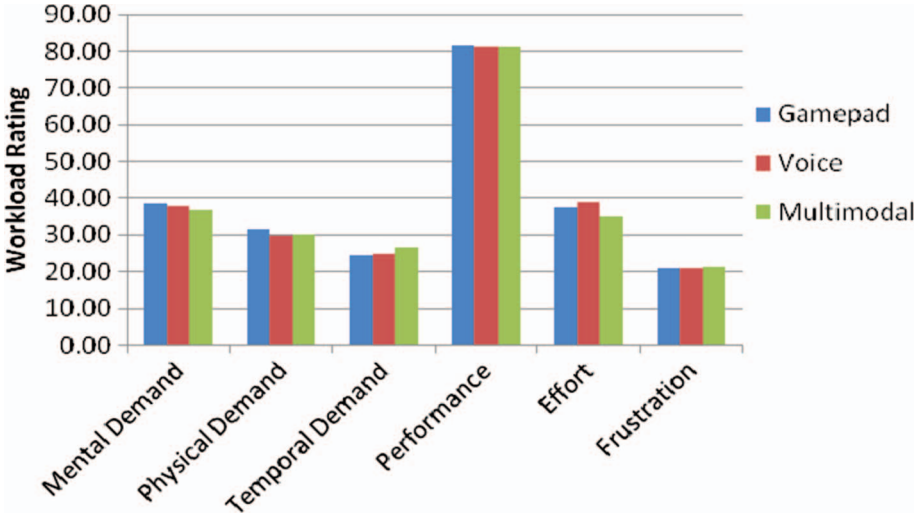

Figure 8. Subjective workload for different conditions.

different modalities in the multimodal condition was compared. Further analysis of the type of command in the final trial showed that participants tended to use the gamepad for controlling the camera (i.e. spatial commands) and speech input for classifying the target (Table 1).

Conclusions

Experiment 1 shows a number of effects that are of interest. None of the measures showed an effect of trial on performance. Inspection of the data suggests that participants quickly learned the task and performance levelled after the first trial for each modality.

There are significant main effects related to control modality. Generally, speech alone led to inferior performance on all objective measures. Participants issued significantly fewer commands when using speech only, which is taken to indicate that speaking commands to control the sensor is more demanding than using gamepad controls. This might be interpreted in terms of incompatibility between spatial operation and speech commands (as discussed in the Introduction). Response to warnings was significantly slower when using speech alone and this was compounded when managing 'air' vehicles. This further implies the demand associated with using spoken commands for the tasks.

The analysis of the Multimodal condition shows that combining speech with manual control leads to a faster response to warnings (than using gamepad or speech alone). Analysis of commands issued in the multimodal condition suggests that, for the majority of participants, speech allowed the separation of target classification commands from vehicle control commands.

Table 1. Combinations of gamepad and speech to enter commands

Control of camera	Acknowledge warnings	Classify target	# Participants
Gamepad	Gamepad	Gamepad	3
Speech	Speech	Speech	1
Gamepad	Gamepad	Speech	6
Gamepad	Gamepad/ Speech	Speech	7

The impact of speaking, on the objective measures, implies some difference in demand on the tasks. While this is not borne out by subjective rating of workload, it could indicate the challenge of formulating a spoken response and the need to schedule this behaviour in the context of performing manual control tasks. From this observation, a possible explanation of the superior performance of participants using multimodal interaction (over speech or gamepad alone) on the secondary (respond to warnings) task could be seen in terms of scheduling. In this case, scheduling involves both the separation of the tasks into 'monitor UV' and 'evaluate target' and the assignment of a specific modality to each task. This allows participants to continue the 'monitor UV' tasks as they manipulate controls during the flight and then to switch attention to speech input when they need to perform the parallel task of target classification.

Experiment 2

Aims

Experiment 1 showed that a multimodal user interface leads to the separation of vehicle control from target classification and this could relate to the scheduling of tasks. In experiment 2, the question is whether the use

of a multimodal interface can support management of several UVs. To this end, participants were required to control one, three or five UVs.

Participants

The same 17 people who participated in experiment 1 took part in experiment 2. There was a gap of 1 week between experiments. It was felt that experiment 1 provided training on the basic tasks involved in sensor control and target classification and experience of using the speech and gamepad controls.

Method

The procedure for this experiment is similar to that for experiment 1. Participants completed a practice run using the multimodal configuration, followed by three trial runs to control one, three or five UVs. Given the relative advantage of the UGV in experiment 1, it was decided to concentrate on UGVs for experiment 2. This was based on the assumption that attending to several UAVs could prove so difficult that participants would miss targets (an assumption that was supported by a short pilot study). Obviously, there is a need to explore the impact of multimodal interaction on airborne UVs but this will be the subject of future work. Commands could be issued using speech or gamepad at the participants' discretion, i.e. multimodal interaction. Each command was logged in terms of time and distance to target. For each trial run, the average of these data was used for analysis. Analysis was conducted using two-way repeated measures ANOVA, which involved modality (speech, gamepad, multimodal) × number of vehicles (one, three, five).

Results

Commands issued

When considering control of one, three or five UVs, there is a main effect of number of vehicles on commands issued [$F(2, 82) = 3.394$, $p = 0.0383$]. The number of commands appears to decrease in proportion to the number of vehicles being controlled. There is also a main effect of modality [$F(1,82) = 29.371$, $p < 0.0001$], with more commands being issued using the gamepad control than using speech, and a significant interaction between number of UVs and modality [$F(2,82) = 3.306$, $p = 0.0416$]. Figure 9 indicates that these effects can be explained by a decrease in number of commands issued using the gamepad as the number of vehicles increase from one vehicle to three or five (there are no differences in number of commands for three and five vehicles).

The number of commands issued using speech does not exhibit so obvious a trend and is more or less constant. Closer inspection of Figure 9 suggests that speech is used to issue around 10% of the total command set. Given that there were always six targets for the participant to classify, it would seem reasonable that speech would be used to issue a small number of commands. This is commensurate with the suggestion in experiment 1 that speech tends to be reserved for target identification and gamepad controls for all other sensor control, which Table 2 shows.

Distance to target

In terms of distance to target, there is a no effect of number of UVs but a main effect of modality [$F(1, 57) = 1.572$, $p = 0.0215$]. This is illustrated by Figure 10.

Time to respond to warnings

In terms of time to respond to warnings, there is no effect of number of UVs but a main effect of modality [$F(1,48) = 11.793$, $p = 0.0012$]. There is a small increase in response time for the gamepad control as the number of targets increases, but a much larger and more obvious increase for speech. Post-hoc testing

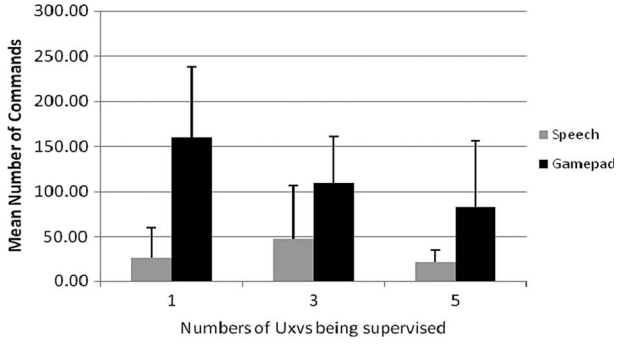

Figure 9. Mean number of commands for each modality.

Table 2. Combinations of gamepad and speech to enter commands

Control of camera	Acknowledge warnings	Classify target	# Participants
Gamepad	Gamepad	Gamepad	3
Speech	Speech	Speech	0
Gamepad	Gamepad	Speech	4
Gamepad	Gamepad/ Speech	Speech	10

indicates a significant difference between one and five vehicles ($p < 0.05$) for speech. This is illustrated by Figure 11.

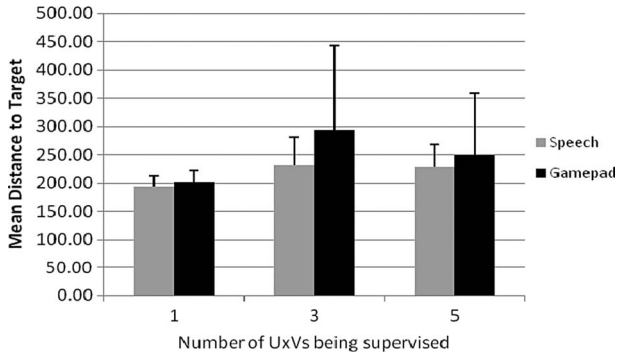

Figure 10. Mean distance to target for each modality.

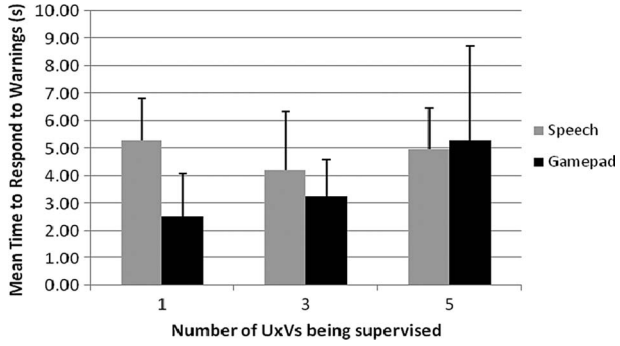

Figure 11. Mean time to respond to warnings for each modality.

Subjective workload

There is a significant main effect of number of UVs on subjective rating of workload [$F(2,252) = 8.395$, $p = 0.0003$]. This is illustrated by Figure 12, which shows how the rating of workload for controlling a single UV is much lower than ratings for three or five UVs. There is less difference in ratings between three and five UVs.

Influence of spatial ability

Performance on the spatial test was used to divide the participants into two groups. The mean score was used to define the split and this gave 10 participants with scores less than the mean and seven participants with scores greater than the mean. While this is a fairly crude break-down, it does create two groups for further analysis. Comparison of the results was performed using the Mann-Whitney U-test (because the data were assumed to be non-parametric). For a single UV, there were no significant differences between groups. However, there was a significant difference between groups for time to respond to warnings with five UVs ($U = 1$, $p < 0.005$) (see Table 3).

Conclusions

Increasing the number of vehicles that a person controls has an impact on subjective workload. This is particularly apparent when moving from managing sensors on a single UV to dealing with three UVs (the effect is less apparent when increasing from three to five

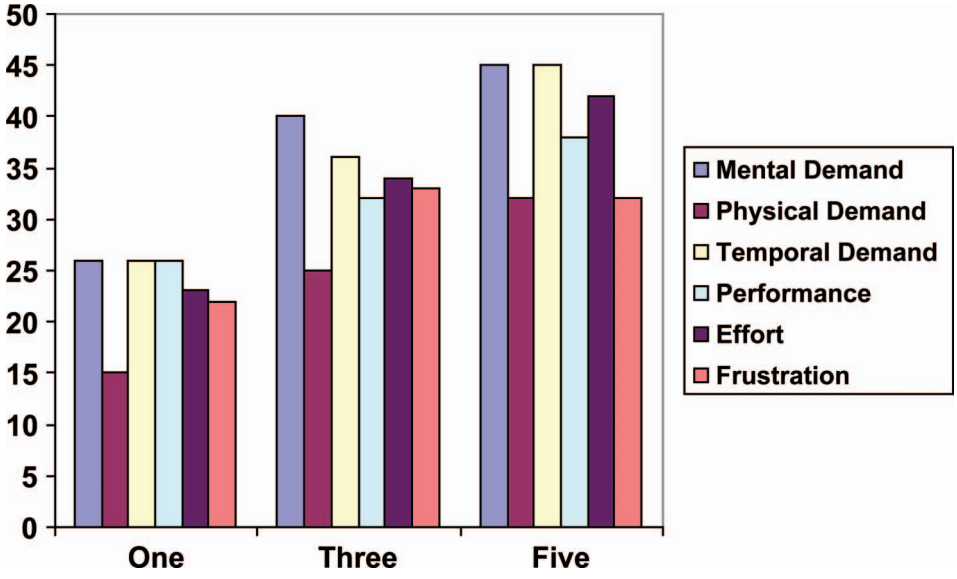

Figure 12. NASA-Task Load Index ratings for numbers of unmanned vehicles.

Table 3. Effect of spatial reasoning for number of UVs (1, 3 or 5)

	1			3			5		
	Cmd	DTT	TW	Cmd	DTT	TW	Cmd	DTT	TW
'Good'	150	192	3	134	235	3	130	236	3
'Poor'	194	194	3	143	231	5	160	219	5
Sig.	n.s.	n.s.	n.s.	n.s.	n.s.	$p = 0.0639$	n.s.	n.s.	$p = 0.0298$

DTT = distance to target; TW = time to respond to warnings.

UVs). Similarly, the increase in UVs being managed has an impact on number of commands issued, particularly when moving from one to three UVs. Thus, in terms of objective measures, increasing the number of vehicles had a bearing on two aspects of performance. First, for the control of the UGVs, there was a decrease in number of commands issued using the gamepad as the number of vehicles increased. Second, in terms of time to respond to warnings, there was a significant increase in response time when using speech when monitoring three or five, rather than one, vehicles.

When considering the relationship between spatial ability and task performance, a significant difference in performance was found when considering time to respond to targets when controlling five vehicles. A possible explanation for this difference is that the participants with low spatial ability found it different to schedule attention between the warnings and the control of many vehicles. In general, experiment 2 also supports the observation that participants prefer to reserve speech control for target classification and use the gamepad for sensor control.

Discussion

Summary of experiments

In experiment 1, participants issued significantly fewer commands when using speech only compared with either gamepad or multimodal control. Target classification was performed much closer to targets under the 'air' than the 'ground' condition, for all modalities. Response to warnings was significantly slower when using speech than other modalities and this was compounded when using speech with 'air' vehicles. There were no effects on workload or of spatial ability. In the multimodal condition, there was a tendency to use the gamepad for sensor manipulation and speech for target classification and response to warnings.

In experiment 2, the number of commands issued decreases as the number of UVs increases. There is also a main effect of modality and an interaction between modality and number of UVs: in particular, number of commands decreases for the gamepad when number of UVs increases from one to three (but not from three to five). There is a main effect of modality on distance to target (as in experiment 1), but not number of UVs. There is also a main effect of modality on time to respond to warnings (as in experiment 1) but not number of UVs. Post-hoc tests show significant difference for speech, when increasing the number of UVs from one to five, in terms of response time. There is a significant increase in subjective workload when the number of UVs increases from one to three, less from three to five. Spatial ability affects time to respond to warnings at five UVs.

The first experiment demonstrated that a multimodal interface, in which participants were able to use a combination of speech and gamepad control, led to superior performance on a secondary 'respond to warnings' task when compared with using either gamepad or speech alone. The main explanation of this advantage of a multimodal interface was that participants opted to issue target classification using speech while controlling the UVs with the gamepad control. This allowed them to divide the tasks of managing sensors from those of classify targets or responding to warnings. Furthermore, performing the task using speech alone led to a demonstrably inferior performance to using the gamepad or multimodal interface, which suggests that care needs to be taken if designs for solely speech-driven interfaces are to be considered. The experiment also contrasted two types of UV and it was shown that the UAV led participants to be closer to the target (in terms of distance over the ground) than the UGV. This indicates different perceptual demands in terms of target size and conspicuity.

The second experiment demonstrated that increasing number of vehicles to control leads to an increase in subjective workload. The number of commands issued, for the gamepad, decreases with an increase in number of vehicles. This indicates a change in strategy to help schedule the response to warnings, i.e. by reducing attention to the task of issuing commands. This is further supported by the increase in time to respond to warnings, using speech, when controlling five vehicles. Again, there are effects of type of vehicle and of modality and, as with experiment 1, it appears that participants prefer to control the UVs using manual control and reserve speech for classifying

targets. In terms of spatial ability, participants with low spatial ability found it difficult to manage five vehicles, as evidenced by the difference in time to respond to warnings.

Comparison of performance across number of UVs

Previous work suggested that there might be an upper limit of four UVs at which performance could begin to deteriorate (Taylor 2006, Cummings *et al*. 2007, Liu *et al*. 2009). Combining performance across the two experiments (Figure 13), one can see that there is an indication that performance changes as more UVs come under the participants' control. This suggests that increasing from one UV to three UVs leads to a change in task demands (both subjective and objective), but there is less increase in demands from three to five, at least for participants with good spatial ability. As noted previously, distance to target can be thought of as a temporal measure as well; the more efficient the performance, the more likely the target will be identified at a greater distance. As the number of UVs increase, so the distance to target decreases. This is true for both modalities, and there is a similar decrement for the two modalities (from around 200 pixels with one UV to around 100 pixels with five UVs). It could be argued that distance to target is less a matter of sensor control than of the participants' ability to schedule attention between the map showing the movement of UV (which influenced judgements related to sensor controls) and the display showing targets, and this ability becomes challenged as workload increases. This might explain why participants with low spatial ability found the task of managing five UVs particularly challenging, as shown by their performance on the secondary task.

Further, time to respond to warnings also differentiates between speech and gamepad when there are five vehicles. An explanation for this could be that time to respond to warnings involves specifying and performing the command action, which could be cognitively demanding when speaking (particularly under higher workload) but (once the button-set was learnt) could be less demanding for the gamepad control.

Multimodal interaction

The experiments demonstrate that using speech alone is costly for this application. Not only is performance on the secondary task lower when using speech, but also participants tended to have their UVs closer to targets for classification. Of course, distance to target (for these simulations) has a strong temporal dimension and it is likely that this is measuring the time taken to prepare and produce a spoken response. Thus, in terms of the scheduling proposal put forward in the Introduction, it is felt that the use of speech for this task imposes a demand on participants that led to a slowing of activity. It is interesting to note that when speech is combined with gamepad, there is a distinct improvement in performance (relative to the use of speech alone).

In broad terms, the number of commands that the participant issues with gamepad shows a decrease. However, the relative number of speech to gamepad control commands remains consistent, which might imply that the strategy employed was consistent across

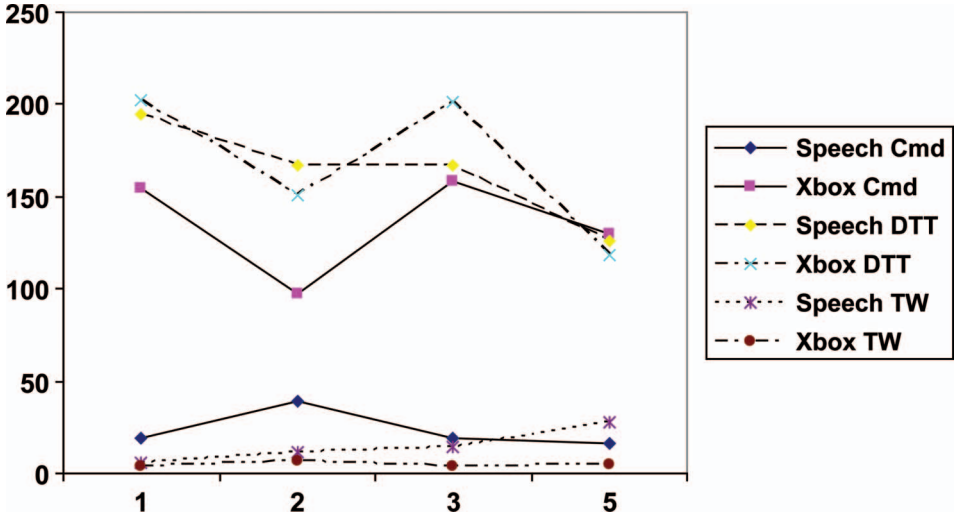

Figure 13. Relative performance for one, two, three and five vehicles, using different modalities, on all measures. Cmd = commands issued.

the trials. That is, as the UV was autonomous, the task of the participant could be summarised in terms of the following routine: check map, when UV approaches target switch to that UV, move camera on to target, speak command, and any warnings to be dealt with when time permitted. This suggests a plausible reason for the distribution of commands between speech and gamepad control. Participants tended to allocate the tasks of controlling sensors (which are spatial tasks) to the gamepad and the tasks of classifying targets and responding to warnings (which are symbolic tasks) to speech. While this allocation was not universal, it was sufficiently strong to suggest a preference. This follows previous work on combining speech with other modalities (Bolt 1980, 1984, Martin 1989, Schmandt *et al*. 1990, Biermann *et al*. 1992, Thorisson *et al*. 1992, Hauptmann and McAvinney 1993, Baber 1997, Cohen *et al*. 1997, 2002, Oviatt *et al*. 1997, 2000). If participants apply a similar strategy across several UVs, then one might expect to see some impairment of performance as more vehicles are managed. In particular, if warnings are dealt with when time permitted, then one might anticipate an increase in time to respond to warnings. This is what one clearly sees in the speech data but it is not so apparent in the gamepad control data – indeed, there seems to be a fairly consistent response over number of vehicles.

Conclusions

Combining speech with gamepad to control sensors on UVs leads to superior performance on a secondary task (implying a reduction in demand) and use of fewer commands (when compared with using either control modality singly). When controlling multiple UVs, performance seems to deteriorate when moving from one to three UVs (which corresponds to previous research). This deterioration is particularly marked for participants with low spatial orientation scores. It is proposed that the findings of these studies have potential implications for user interface design for future UV workstations and for selection of UV operators.

Acknowledgements

This work is supported by a grant from the Human Factors Integration Defence Technology Centre funded by the Human Dimension and Medical Sciences Domain of the UK Ministry of Defence Scientific Research Programme.

Notes

1. The experiment has been designed to emphasise the control of sensors rather than the interpretation or analysis of imagery. It is suggested that the primary difference between an experienced UV payload operator and the participants in the present study would relate more to the response to targets than the direct control of sensors. For example, observations of payload operators and discussion with subject matter experts have emphasised the ability to anticipate target behaviour, to draw on multiple sources of knowledge to interpret the target and to accurately describe the target to colleagues. Care was taken to remove these demands from this simplified, abstracted version of the task.
2. The 11 commands, which could be issued using gamepad controls or spoken commands, were 'up', 'down', 'left', 'right', 'centre', 'zoom_in', 'zoom_out', 'acknowledge', 'odd', 'even', 'switch'.
3. http://spatiallearning.org/resource-info/Spatial_Ability_Tests/PTSOT.pdf

References

Baber, C., 1997. *Beyond the desktop: Designing and using interaction devices*. San Diego, CA: Academic Press.

Baber, C. and Mellor, B., 2001. Using critical path analysis to model multi-modal human-computer interaction. *International Journal of Human-Computer Studies*, 54, 613–636.

Baddeley, A.D., 1998. The central executive: a concept and some misconceptions. *Journal of the International Neuropsychological Society*, 4, 523–526.

Baddeley, A.D., 2007. *Working memory, thought and action*. Oxford: Oxford University Press.

Biermann, A.W., Fineman, L., and Heidlage, J.F., 1992. A voice and touch driven natural language editor and its performance. *International Journal of Man-Machine Studies*, 37, 1–21.

Bolt, R.A., 1980. Put-that-there: voice and gesture at the graphics interface. *Computer Graphics*, 14, 262–270.

Bolt, R.A., 1984. *The human interface: Where people and computers meet*. Boston: Lifelong Learning Publications.

Byers, J.C., Bittner, A.C., and Hill, S.G., 1989. Traditional and raw Task Load Index (TLX) correlations: are paired comparisons necessary? *In*: A. Mital, ed. *Advances in industrial ergonomics and safety 1*. London: Taylor and Francis, 481–485.

Chen, J.Y.C., 2008. UAV-guided navigation for ground robot tele-operations in a military reconnaissance environment. *Ergonomics*, 53, 940–950.

Chen, J.Y.C., 2009. Concurrent performance of military and robotics tasks and effects of cueing in a simulated multi-tasking environment. *Presence*, 18, 1–15.

Chen, J.Y.C., Haas, E.C., and Barnes, M.J., 2007. Human performance issues and user interface design for tele-operated robots. *IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews*, 37, 1231–1245.

Cohen, P.R., *et al*., 1997. Quickset: multimodal interaction for distributed applications. *5th ACM conference on Multimedia*. New York: ACM, 31–40.

Cohen, P.R., *et al*., 2002. Multimodal interaction for 2D and 3D environments. *IEEE Computer Graphics and Applications*, 19, 10–13.

Cummings, M.L., *et al*., 2007. Automation architectures for single operator, multiple UAV command and control. *The International C2 Journal*, 1, 1–24.

Dixon, S.R. and Wickens, C.D., 2003. Control of multiple-UAVs: a workload analysis. *12th international symposium on aviation psychology*. Dayton, OH: Association for Aviation Psychology.

Donmez, B., Cummings, M.L., and Graham, H.D., 2009. Auditory decision aiding in supervisory control of multiple unmanned aerial vehicles. *Human Factors*, 51, 718–729.

Draper, M., *et al*., 2003. Manual versus speech input for unmanned aerial vehicle control station operations. *In*: *Proceedings of the 47th annual meeting of the Human Factors and Ergonomics Society*, Santa Monica, CA: HFES, 109–113.

Finn, A. and Scheding, S., 2010. *Developments and challenges for autonomous unmanned vehicles: a compendium*. Berlin: Springer-Verlag.

Hart, S.G. and Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *In*: P.A. Hancock and N. Meshkati, eds. *Human mental workload*. Amsterdam: Elsevier, 139–183.

Hauptmann, A.G. and McAvinney, P., 1993. Gestures with speech for graphic manipulation. *International Journal of Man-Machine Studies*, 38, 231–249.

Hegarty, M. and Waller, D., 2004. Dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32, 175–191.

Helleberg, J., Wickens, C.D., and Goh, J., 2003. Traffic and data link displays: auditory? Visual? Or redundant? A visual scanning analysis. *12th international symposium on aviation psychology*. Dayton, OH: Association for Aviation Psychology.

Kahneman, D., 1973. *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kozhevnikov, M. and Hegarty, M., 2001. A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory and Cognition*, 29, 745–756.

Liu, D., Wasson, R., and Vincenzi, D.A., 2009. Effects of system automation management strategies and multi-mission operator-to-vehicle ratio on operator performance in UAV systems. *Journal of Intelligent Robotics Systems*, 54, 795–810.

Luck, J.P., *et al*., 2006. An investigation of real world control of robotic assets under communication. *In*: *Proceedings of 2006 ACM conference on human-robot interaction*. New York: ACM Press, 202–209.

Martin, G.L., 1989. The utility of speech input for user computer interfaces. *International Journal of Man-Machine Studies*, 30, 355–375.

Maza, I., *et al*., 2009. Multimodal interface technologies for UAV ground control stations:a comparative analysis. *Journal of Intelligent and Robotic Systems*, 57, 371–391.

Moroney, W.F., *et al*., 1995. Some measurement and methodological considerations in the application of subjective workload measurement techniques. *The International Journal of Aviation Psychology*, 5, 87–106.

Nygren, T.E., 1991. Psychometric properties of subjective workload measurement techniques: implications for their use in the assessment of perceived mental workload. *Human Factors*, 33, 17–33.

Oviatt, S., Deangelli, A., and Kuhn, K., 1997. *Integration and synchronization of input modes during multimodal human-computer interaction*. *CHI'97*. New York: ACM, 415–422.

Oviatt, S.L., *et al*., 2000. Designing the user interface for multimodal speech and gesture applications: state-of-the-artsystems and research directions. *Human-Computer Interaction*, 15, 263–322.

Schmandt, C., *et al*., 1990. Observations of using speech input for window management. *In*: D. Diaper, D. Gilmore, C. Cockton, and B. Shackel, eds. *Interact'90*. Amsterdam: IOS, 787–793.

Taylor, R.M., 2006. Human automation integration for supervisory control of UAVs. *In*: *Virtual media for military applications*, meeting proceedings Rto-Mp-Hfm-136, Paper 12. Neuilly-Sur-Seine, France: RTO.

Thorisson, K.R., Koons, D.R., and Bolt, R.A., 1992. Multimodal natural dialogue. *CHI'92*. New York: ACM, 653–654.

Trouvain, B. and Schlick, C.M., 2007. A comparative study of multimodal displays for multirobot supervisory control. *In*: D. Harris, ed. *Engineering psychology and cognitive ergonomics*. Berlin: Springer-Verlag, 184–193.

Wickens, C.D., Dixon, S., and Chang, D., 2003a. *Using interference models to predict performance in a multiple-task UAV environment*. Technical Report AHFD-03–9/Maad-03–1. Savoy, IL: Aviation Human Factors Division, Institute of Aviation, University of Illinois at Urbana-Champaign.

Wickens, C., Wickens, D., and Liu, Y., 1988. Codes and modalities in multiple resources: a success and a qualification. *Human Factors*, 30, 599–616.

Wickens, C.D. and McCarley, J., 2008. *Applied attention theory*. Boca-Raton, FL: Taylor and Francis.

Wickens, C.D., Sandry, D., and Vidulich, M., 1983. Compatibility and resource competition between modalities of input, output, and central processing. *Human Factors*, 25, 227–248.

Wickens, C.D., *et al*., 2003b. Attentional models of multi-task pilot performance using advanced display technology. *Human Factors*, 45, 360–380.